Tartan Data Science Cup 2017

February 4, 2017

Introduction

Using slightly altered data from Lending Club, the competition this year investigated whether certain features of loans correspond to the likelihood that the loan is paid off. We were given two data sets, one with preserved entries and another corrupted set with missing loan status information. Our purpose in this competition was to predict the loan status outcome for the corrupted data set, which fell into either Group 0 ("Good") or Group 1 ("Bad"). We predicted loan status outcomes by using a decision tree trained on the preserved set and running a second algorithm that considered the weighting of significant factors.

Methods

Data Scrubbing

We conducted our analysis in Python using the `pandas`, `NumPy`, and `sklearn` packages. In order to visualize and process the data sets, we needed to convert all of the data into numerical values to quantify each category. For example, we converted `sub_grade` values, which ranged from `['A1', 'A2', ..., 'G5']`, to integers `[0, 1, 2, ...]`. For values that were missing, we set them to values such as `-1` so that we could catch them later on in our classification algorithm.

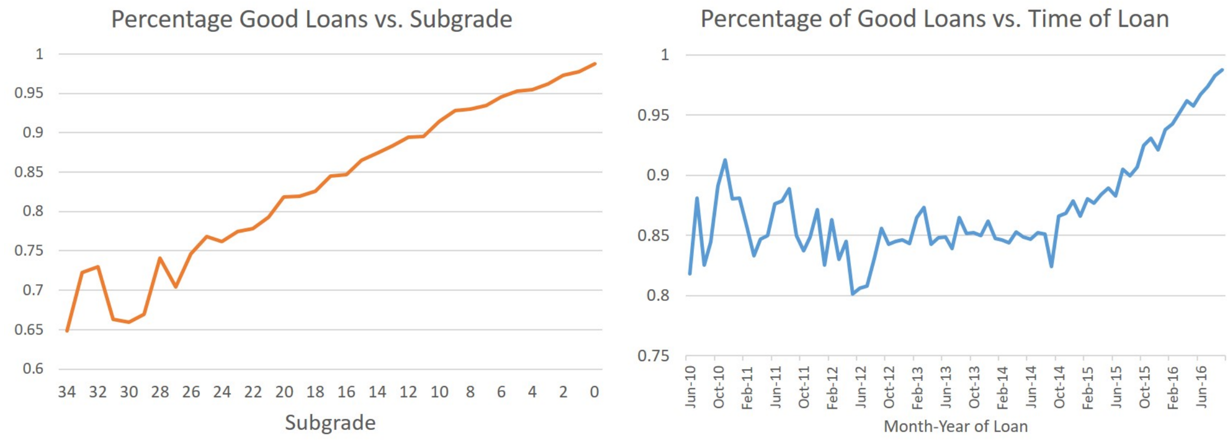
Tree

We used a supervised learning method for training our decision tree. We tried many different combinations of estimators and training set compositions in order to maximize the accuracy of our tree's prediction. One of the biggest challenges we ran into with working with the data set is how heavily skewed it was towards good loan status outcomes versus bad loan status outcomes. In order to resolve this bias, we downscaled our data set by filtering out the total number of good people so our training set was in a 80 : 20 ratio of good to bad loans.

We first compiled a tentative list of estimators by correlating parameters with loan status outcomes. For example, as seen in Figure 1.a), loans that were recently taken out were very likely to have a good loan status since they were too recent to incur fines. After running our tree with this initial list of estimators, we ran into the problem of over-fitting the tree which lead to false negatives that were impacting our accuracy.

In response, we isolated each estimator we were interested in and calculated its Gini importance value; our final decision tree's estimators were comprised of only those with the highest Gini importance value. The final estimators that were used by our decision tree were loan amount, loan subgrade, annual income, issue date of the loan, debt income ratio (DTI), number of open credit accounts, number of total credit accounts, and house ownership status.

We also decided to split our data based on the observation that loan status was strongly correlated with loan grade. As seen in Figure 1.b), lower subgrades (correlating to higher quality loans such as A, B grade loans) were associated with very high percentages of good loan statuses. So, our predictor function would first consider whether the grade status of the loan was high and then appropriately predict a good loan status, and then predict the remaining loans using our decision tree.



(a) The percentage of good loans increases with better subgrades

(b) More recent loans are more often considered "good", probably because many of them are still current

Figure 1: Graphs with significant weighting on status of loan

Key Results

Score and Output

The competition uses Brier Score to compare submissions. The formula for Brier score is

$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2. \tag{1}$$

where f_t is the predicted output and o_t is the actual output.

We computed a Brier Score on our validation data (which was the last around 10,000 entries of the given training data) to fine tune our submission. The final Brier score for the validation data was 0.117. In our output, 1745 (17%) out of the 10,000 loans were predicted as bad, which suggests a bias towards labeling good loans as bad rather than bad loans as good (false negatives) since the actual ratio should have been around 10%.

Conclusion

We successfully implemented a full stack approach to visualizing, analyzing, and predicting whether or not certain features point to the possibility of loaners paying off their debt. Our results indicate that certain factors, such as subgrade, loan amount, and income debt ratio have a significant bearing on the final result of a loan. Our model was also very accurate in catching determining loans that would have a negative outcome. Some future improvements on our model include accounting for overfitting more, including more features, implementing a way to combine the decision tree with other prediction algorithms, and finding a way to avoid false negatives.